Artificial & Computational Intelligence /Published Online: March 2021
https://acors.org/ijacoi/VOL2_ISSUE1_03.pdf

Artificial & Computational Intelligence

Resampling Imbalance Class to Analyze Performance of Machine Learning Algorithms for Classification

Chitra Desai
Professor and Head, Faculty of Computational Science
chitragdesai@gmail.com
National Defence Academy, Pune, Maharashtra, India

Abstract

The performance of machine learning algorithm depends on several factors mainly including quality and quantity of the data, sampling techniques, feature engineering, feature selection, algorithm tuning, hyper parameters to list the few. Along with these, the performance of model for a given dataset in machine learning is also attributed to achieving balance between underfitting and overfitting. Further when machine learning algorithm for classification is applied to a dataset, the metrics – precision, recall, accuracy and F1 score are useful in gaining insight into the predictive capability of the model. Given a data set with imbalance class to be trained on machine learning model for classification, can introduce learning problem. The insight into these problems can be gained using these metrics and particularly is more useful when the data is related to medical diagnosis. To resolve the problem of imbalance and improve on predictive capabilities of the machine learning model, imbalance data handling techniques with other data pre-processing steps can be found useful. This paper aims at applying techniques to balance data and observing the impact on predictive capabilities of the machine learning models. Further, ensemble machine learning model is applied on the data set to observe the impact on model accuracy.

Keywords: Imbalance class, Logistic Regression, Decision Tree, KNN, Oversampling, Undersampling, Nearmiss

1. Introduction

Classification using machine learning algorithms in medical diagnosis [1] is gaining popularity with the volume of data being available to train machine learning algorithms. The large-scale real data that is available for training the machine learning algorithms however poses the challenges of imbalance class distribution. This is because the performance of the machine learning algorithm is impacted by the imbalance class [2]. There are several other factors that impacts the performance of a learning algorithm [3]. Although factors like redundant attributes, noise, outliers, data transformation are commonly delt while pre-processing the data, often, the skewed distribution of target variable goes unattended. This eventually brings in the biasness into the model due to majority class and there is high probability that the minority class may get misclassified compared to majority class. It is also likely that minority class may get treated as noise and get overlooked. In order to handle imbalance class in a dataset, resampling can be applied to either increase the number of minority class or decrease the number of majority class. By treating the imbalance class of the training data before providing it as an input for machine learning algorithm it is possible to improve the accuracy of machine learning algorithms for classification.

For classification itself we find there are different machine learning models available. The predictive capability of each of the model vary according to the modelling techniques used, difference in population, the initial seed etc apart from imbalance class. One way to improvise on stability and predictive capability of the model is to combine individual models in machine learning to achieve ensemble modelling [4][5].

This paper aims at applying techniques of imbalance data handling – oversampling, undersampling and nearmiss to the data set. The result with and without resampling are compared to observe the impact on predictive capabilities of the machine learning models. The three machine learning algorithms used are logistic regression (LR), decision tree (DT) and K- nearest neighbor (KNN). Further, ensemble machine learning model is applied on the data set to observe the impact on model accuracy to the balanced dataset.

2. Methodology

The methodology adopted here is to initially observe the predictive capability of the three-machine learning (ML) models LR, DT and KNN on the original data set. Identify the class distribution of target variable and accordingly apply resampling techniques to balance the data set. Three techniques used for balancing the data set are – oversampling, undersampling and nearmiss. After balancing the dataset, with each of the three techniques, the three ML models are trained separately to observe the impact of each of the balancing technique on each of the three ML algorithms. Also, ensemble modelling on balanced dataset is applied to identify the impact on model accuracy.

3. Data

In this paper the impact of imbalance data handling techniques on performance of three machine learning algorithms is verified using the orthopaedic patient data set [6]. The original data set consist of 310 records (rows) and 7 columns. The dataset is imbalanced as the classification categories are not equally represented. It consists of 60 patients from class Hernia, 100 patients from class Normal and 150 patients from class Spondylolisthesis. The data is pre-processed to ensure that there are no missing values and no duplicates.

4. Machine Learning Algorithms

The three algorithms used to classify the target variable in this paper are, Logistic Regression, Decision Tree and K-NN. The pre-processing on data and test accuracy using these three models is already as shown in [7]. The experiments were conducted initially on the original data set to predict the test accuracy using Logistic regression, Decision tree and K-NN algorithm. Another performance metric that can be used is error rate. Error rate = 1- Accuracy. The test accuracy and error rate results for the three models are as shown in Table 1. The LR model shows better test accuracy compared to DT and K-NN. As the class distribution is uneven to gain better insight, F1 score is computed for the three models. Table 2 presents the precision, recall and F1 score for the three models for each class.

Table 1 Test Accuracy Score for LR, DT and KNN

Model	Test Accuracy Score	Error rate = 1-Accuracy
Logistic Regression (LR)	0.838710	0.1613
Decision Tree (DT)	0.763441	0.2366
K-NN	0.784946	0.2151

Table 2 Precision, Recall and F1 Score before resampling

Model	Class	Precision	Recall	F1-Score
LR	Hernia	0.56	0.5	0.53
DT		0.5	0.44	0.47
KNN		0.56	0.5	0.53
LR	Normal	0.62	0.67	0.64
DT		0.56	0.62	0.59
KNN		0.62	0.67	0.64
LR	Spondylolisthesis	0.94	0.94	0.94
DT		0.96	0.94	0.95
KNN		0.94	0.94	0.94

It is observed here in table 2 that F1 score for class-Spondylolisthesis is comparatively far better than other two classes in all the three models because the number of observations for the class – spondylolisthesis is 150 and we therefore observe the biasness in model prediction. This biasness is observed due to data imbalance.

5. Imbalance Data Handling Techniques

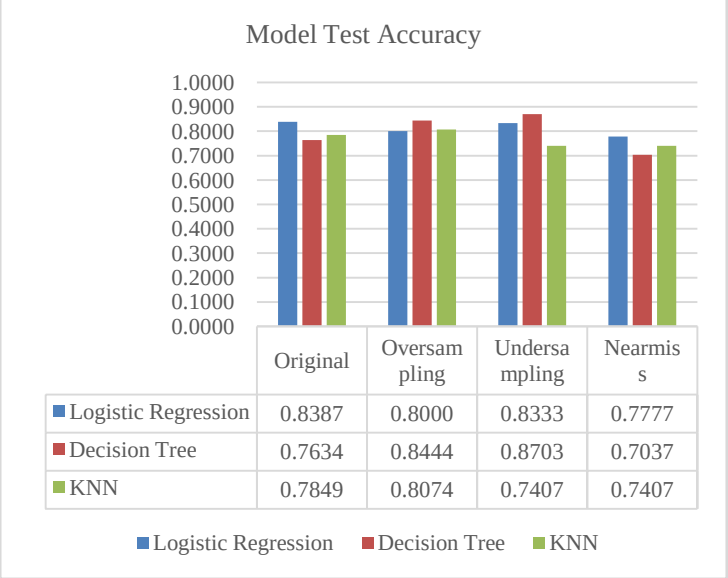
The real-life data samples often show high class of imbalance, which needs to be determined before exposing the training data to model in machine learning algorithms. There are several techniques for balancing data [8]. In this paper we have experimented with three techniques- oversampling, undersampling and nearmiss.

In oversampling random replication of minority class is done to balance the target class. Oversampling can lead to overfitting [9][10] and also it is possible that if the difference between the majority class and multiple minority classes is too high, the replication of minority classes may demand more computational resources. Synthetic Minority Oversampling Technique (SMOTE) [9] is applied on the original data set and the resultant data set consist of 450 records i.e., 150 (33.33%) records of each of the three classes.

In undersampling and nearmiss [11] the original data set is reduced to 180 records i.e., 60(33.33%) records of each of the three classes. In undersampling the majority class is under sampled by randomly selecting samples with replacement. With replacement set to true bootstrapping is achieved [12]. In nearmiss under sampling is done by applying the heuristic rules based on nearest neighbor algorithm. Here under sampling using nearmiss is applied with 7 nearest neighbors.

6. Experimental Results

It is observed that when oversampling of the data is done and then fitted to the models – Decision tree and KNN showed improvement in prediction of accuracy compared to fitting to the original dataset. However, using nearmiss the performance is seen to be degenerated in case of all the three models. Also, using undersampling we see improvement in predictive capability of decision tree algorithm compared to that of oversampling. The results are as shown in figure 1.



	Original	Oversampling	Undersampling	Nearmiss
Logistic Regression	0.8387	0.8000	0.8333	0.7777
Decision Tree	0.7634	0.8444	0.8703	0.7037
KNN	0.7849	0.8074	0.7407	0.7407

Figure 1 Figure 1 Test Accuracy after Resampling

Another experiment was performed where ensemble of ML algorithms with three different base classifiers were applied on resampled data. The results are as shown in figure 2. It is observed that nearmiss in case of all the three base classifier showed degenerating performance. Base classifier with logistic regression also showed comparatively low performance using resampling techniques. The performance with DT as base classifier in oversampling and random undersampling showed improvement compared to the original data set.

Table 3 shows error rate for three ML algorithms applied on original data and balanced data using three different techniques. It is observed that when logistic regression is applied to balanced data obtained using the three techniques, the error rate has increased compared to original data, thereby, reducing the predictive accuracy. It is also observed that using nearmiss techniques when the data was balanced and given as input to ML algorithms, all the three algorithm – LR, DT and KNN showed increase in error rate.

From the results it is concluded that when data is balanced using oversampling and undersampling techniques, the decision tree algorithm shows decrease in error rate. Further, when ensemble of machine learning algorithm with base classifier LR, DT and KNN is applied to original as well as balanced data obtained using the above three techniques, it is

observed that error rate is lowest for balanced data using undersampling techniques with base classifier as DT.

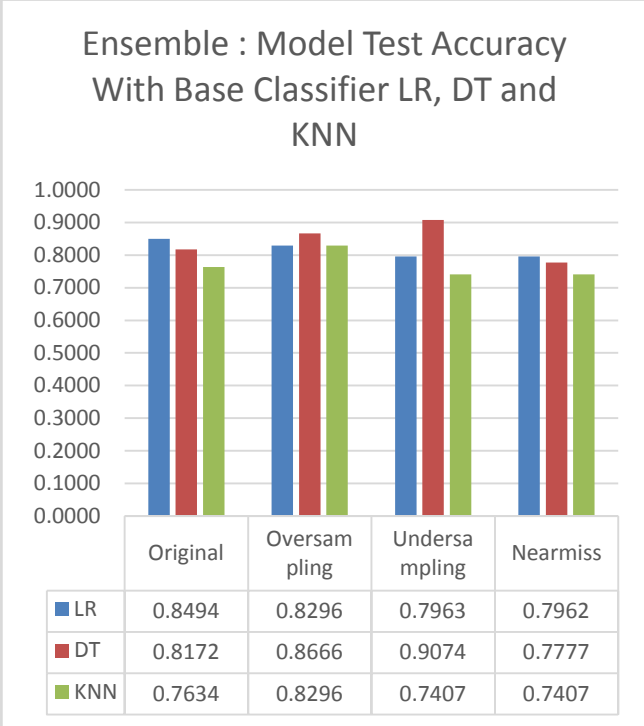


Figure 2 Test Accuracy after Resampling and Ensemble

Table 3 Error rate for ML Algorithms without Ensemble

	Original	Oversampling	Undersampling	Nearmiss
LR	0.1613	0.2000	0.1667	0.2223
DT	0.2366	0.1556	0.1297	0.2963
Knn	0.2151	0.1926	0.2593	0.2593

Table 4 Error rate for Ensemble Model with Base Classifier LR, DT and KNN

	Original	Oversampling	Undersampling	Nearmiss
LR	0.1506	0.1704	0.1704	0.2038
DT	0.1828	0.1334	0.1334	0.2223
Knn	0.2366	0.1704	0.1704	0.2593

7. Conclusion

There are different machine learning algorithms which can be applied to solve classification problem. These models when applied to a data set may vary in accuracy prediction depending upon the flexibility of the model or the restrictions imposed using a particular approach. There are therefore several factors that impact the performance of a model and one such reason can be the imbalanced data set. In this paper, we have used three different techniques – versampling, undersampling and nearmiss to balance the data set and observe its impact on accuracy prediction on three different machine learning algorithms. These three algorithms use three different approaches they are – logistic regression, KNN and decision tree. These algorithms are applied to data set which has multiclass target variable and has imbalanced class distribution. It is observed that logistic regression predicted highest model accuracy of 83.87% without ensemble and 84.94% with ensemble on original data set. The data was balanced using all the three techniques and the models were trained with each of the balanced data set. It was observed that using oversampling DT and KNN outperformed compared to original data set. It was observed that decision tree algorithm using random undersampling gave highest model test accuracy of 87.03%. Further, when ensemble of machine learning model was done, it is observed that Decision tree with undersampling resulted to higher model test accuracy of 90.74%, here the base classifier used was decision tree. Thus, we observe that different models with different imbalance data handling techniques on the same data set may result in varied performance evaluation.

References

[1] Igor Kononenko, Machine Learning For Medical Diagnosis: History, State Of The Art And Perspective, Artificial Intelligence in Medicine, Volume 23, Issue 1, August 2001, Pages 89-109.

[2] Haibo He; Edwardo A. Garcia, Learning from Imbalanced Data, IEEE Transactions on Knowledge and Data Engineering ,Volume: 21, Issue: 9, Sept. 2009

[3] MariamMoustafa Reda , Dr Mohammad Nassef , Dr Akram Salah, Factors Affecting Classification Algorithms Recommendation: A Survey, 8th International Conference on Soft Computing, Artificial Intelligence and Applications (SAI 2019), June 29-30, 2019, Copenhagen, Denmark

[4] Brown, G. Ensemble Learning. In Encyclopedia of Machine Learning; Springer: Boston, MA, USA, 2010; Volume 312

[5] Polikar, R. Ensemble learning. In Ensemble Machine Learning; Springer: Boston, MA, USA, 2012; pp. 1–34.

[6] Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science

[7] Chitra Desai, Classification of Orthopaedic Patients Based On Biochemical Features Using Machine Learning Algorithms, Journal of Critical Reviews 7.18 (2020), 3823-3828. Print. doi:10.31838/jcr.07.18.472

[8] Sotiris Kotsiantis, Dimitris Kanellopoulos, Panayiotis Pintelas, Handling Imbalanced Datasets: A Review, GESTS International Transactions on Computer Science and Engineering, Vol.30, 2006

[9] N. V. Chawla, L. O. Hall, K. W. Bowyer, and W. P. Kegelmeyer. SMOTE: Synthetic Minority Oversampling TEchnique. Journal of Artificial Intelligence Research, 16:321- 357, 2002

[10] M. Kubat and S. Matwin. Addressing The Curse Of Imbalanced Training Sets: One Sided Selection. In Proceedings of the Fourteenth International Conference on Machine Learning, pages 179-186, Nashville, Tennesse, 1997. Morgan Kaufmann.

[11] https://imbalanced-learn.org/stable/under_sampling.html Accessed 20 Apr2021

[12] David Ten, Imbalanced Dataset with Imbalance Learn, https://xang1234.github.io/louvain/ August 2018. Accessed 22 Apr2021